Some Statistical Issues in Forensic DNA Profiling

by

Seymour Geisser
University of Minnesota

Technical Report No. 609
August 1995

Some Statistical Issues in Forensic DNA Profiling

Seymour Geisser*

University of Minnesota

Abstract

We present a discussion of some of the flaws and problems associated with the statistical methodology used by laboratories in DNA profiling and "matching".

*Work supported in part by NIH Grant GM 25271 and the Lady Davis Trust

1 Introduction

A rich source of highly polymorphic genetic markers based on recombinant DNA technology was described by Botstein et al. (1980) that would lead to a human genetic linkage map. These markers are called restriction fragment length polymorphisms (RFLP). A form of RFLP is generated by the presence of variable number tandem repeats (VNTR). A variation of the VNTR developed by Jeffreys et al. (1985a,b) used probes that recognized multiple loci was designated as DNA fingerprinting. Jeffreys then suggested its use for forensic identification. A description of its first "successful" use in a murder case in England known as the "Pitchfork" case became the subject of a popular book by Wambaugh (1989). Actually in this case it was used to exonerate one suspect and its potential use was sufficient to induce the murderer to confess without actually being DNA "fingerprinted". Starting in about 1988, in the United States the technology was mainly to use single locus probes for DNA profiling in forensic identification. The main laboratories were Lifecodes, Cellmark and slightly later the Federal Bureau of Investigation. Currently there are many state and county laboratories doing DNA profiling.

At each VNTR locus (a genetic location on a chromosome) an individual's genotype is expressed by two alleles (genetic expressions), one inherited from one's mother and the other from one's father. These alleles are basically a discrete number of tandem repeats that may vary greatly anywhere from 30 to well over 100, depending on the locus. Current technology probes a locus using electrophoresis to obtain a sizing of the fragment lengths (alleles). These alleles are sometimes referred to as "junk" DNA since they are presumed to be from noncoding regions of DNA and possibly not affected by selective forces. However, this is an unproven presumption rather than a fact.

A profile for an individual is then developed by probing several loci and reporting the two alleles at each locus. When two alleles are the same at a locus this is termed homozygotic and when different, heterozygotic.

The "revolutionary" aspects over previous blood groups profiled for forensic identification are the very large number of possible alleles at a locus and the fact that the profile can be obtained from various human tissues, including blood, semen and hair.

However, there are some problems associated with RFLP-VNTR profiling. The elec-

trophoretic method does not yield a precise resolution of the VNTR values associated with the bands at the locus but is subject to measurement error unlike many of the older blood groups. Further, for the probes used, the number of different possible alleles and the interval in sizing between them is unknown. This requires statistical criteria for determining whether the two allelic band values can be considered similar or not. This is necessary to determine whether a crime scene profile "matches" either a suspect or victim. The second problem facing the forensicist is the so-called rarity of a profile consisting of several probes.

We shall present a critique of the statistical methods used by forensic laboratories.

2 The Initial Comparison (or "Match")

The major forensic laboratories—Cellmark, Lifecodes and the Federal Bureau of Investigation—all have somewhat different statistical approaches to the initial comparison because of the error in measuring a band value induced by the electrophoretic process. Lifecodes generally assumes that 2 band values match (really are similar) if the observed difference between them is not more than 1.8% of their average. According to Lifecodes, Baird et al. (1986), the standard deviation of the difference between two alleles measured on the same gel is .6% of the band size, hence the tolerance is 3 standard deviations. If the misclassification of the match/nonmatch procedure were due only to normally distributed measurement error and not to a host of other possibilities that can occur in a laboratory, then it is clear that theoretically the false exclusion rate (asserting that two bands do not match when in fact they do) is .0027 or 1/370. It is not possible to easily calculate the false inclusion rate (assuming that two bands match when in fact they don't). This can best be done by proficiency tests that are external and blind and of sufficient size to estimate the false exclusion rate. Such tests would include not only measurement errors but other kinds of errors as well. Apparently the only external test that Lifecodes has been subjected to was run by the California Association of Crime Laboratory Directors (CACLD) on a profile of 4 probes. In a total of 100 samples sent, 85 were analyzed without error. The others were inconclusive.

Cellmark changed from using standard deviations to a system of resolution limits

on a band which it defines as ± 1 millimeter on the autoradiograph. If the two bands from a single locus probe are no more than a resolution limit apart they are declared to match. A resolution limit is also expressed as a varying percent of band size, varying from 1.15% to 5.15% of the band size. The larger the band size the larger the percent of the band size. It would appear from some studies of the variation of repeat measurements conducted by Cellmark that the one resolution tolerance on the difference between 2 bands translates approximately to between 3 and 4 standard deviations depending on the band size. Again the false exclusion rate due only to normally distributed measurement error is approximately between 1/370 and 1/15800 for a match between two bands.

As indicated before, the false inclusion rate can only be measured by external proficiency testing. This was also done by the CACLD and Cellmark's false inclusion rate over approximately 100 trials was 1/50.

The FBI asserts that 2 bands match if their difference is no larger than 5% of their average size, Budowle et al. (1991). A study of repeated measurements on pairs from the same individual on the same gel yields a standard deviation of the difference to be .744% of the band size. Hence the FBI's 5% tolerance translates into about 6.7 standard deviations. The false exclusion rate based solely on normally distributed measurement error on a match of two bands is less than 10^{-10}. No blind external proficiency tests have been reported for the FBI. The FBI's claim, Budowle et al. (1991), that this favors a defendant seems preposterous since even if 8 bands of 4 probes constituting a profile were independent, the theoretical false exclusion rate of the profile would still be less than 10^{-8}. However, in an examination of an unpublished study by the FBI of 225 agent-trainees DNA profiled on two occasions, it was ascertained that many did not match themselves by the FBI's own standards, United States vs. Yee, Thompson (1993). The FBI claimed that different conditions existed for the two occasions.

3 A Statistical Test for the Initial Match

The theoretical false exclusion rates due only to measurement error depend on the reasonable assumption that two bands are the same but differ only on normally distributed measurement error. For a probe consisting of 2 pairs of heterozygotic measurements

(X_1, Y_1) and (X_2, Y_2) that require comparison, the assumption of bivariate normality of

$$Z = (X_1 - X_2), \quad W = (Y_1 - Y_2)$$

with highly correlated measurement error appears to be standard model for this situation, c.f. Berry et al. (1992) with estimate $\hat{\rho} = .9$. Given this situation, then under the null hypothesis Z and W represent measurement error and

$$(Z, W) \sim N(\mathbf{O}, \Sigma)$$

$$\Sigma = \begin{pmatrix} \sigma^2 & \rho\sigma\tau \\ \rho\sigma\tau & \tau^2 \end{pmatrix}$$

where

$$\rho \doteq .9, \quad \sigma \doteq .0074\left(\frac{x_1 + x_2}{2}\right), \quad \tau = .0074\left(\frac{y_1 + y_2}{2}\right).$$

Hence under the hypothesis of a match

$$Y = \frac{1}{(.0074)^2(1 - .9^2)}\left[\left(\frac{X_1 - X_2}{\overline{x}}\right)^2 - \frac{2 \times .9(X_1 - X_2)(Y_1 - Y_2)}{\overline{xy}} + \left(\frac{Y_1 - Y_2}{\overline{y}}\right)^2\right]$$

is approximately χ^2 with 2 degrees of freedom. At a given α then one would reject a match if $Y \geq y_\alpha$ where y_α is such that $\Pr[\chi^2_2 \geq y_\alpha] = \alpha$. This should result in a better criterion for a match than treating each Z and W separately, and resolve to a large degree the problem of bandshifting. Thus a simple approximate significance test is available for testing whether there is a match for a probe, that allows for the fact that it is known that measurement error is highly correlated and can result in bandshifting. The forensic laboratories disregard this correlated error by applying their criteria to the components of the pairs individually rather than jointly. At any rate, the plausibility of a match is not graded but decided yes or no.

4 The Relative Frequency of a Profile

Once a match is asserted, the next step for the forensicist is to determine how rare a profile is in a population or the prediction of the fraction of individuals that match the crime scene profile. This is framed as the probability that an unrelated individual chosen at random from an "appropriate" population will also match the crime scene profile. The

logic for this calculation can be described as follows. Let A = accused, C = culprit, M = asserted match of crime scene profile and accused profile, or profile of victim and profile of evidence found on the accused. Now it is tacitly assumed that $\Pr[M|A = C] = 1$. Then adopting the common (though false) simplifying assumptions here, the likelihood ratio is

$$\frac{\Pr[M|A = C]}{\Pr[M|A \neq C]} = \frac{1}{\Pr[M|A \neq C]}. \tag{4.1}$$

The denominator in (4.1) is interpreted as the chance that a individual chosen at random from an appropriate population will also be declared a match. Then this is a quantity that the forensic laboratories will estimate and submit in their report to the court. Prosecutors, judges, jury etc. will generally interpret this as the odds that the accused is guilty. This is often termed the "prosecutors" fallacy or more generally transposing the conditional, c.f. Thompson and Schumann (1987). This is, of course, if correctly computed, a likelihood ratio and not the posterior odds of guilt. To turn this into posterior odds requires prior odds that the accused is the culprit, which then would result in the posterior odds ratio as

$$\frac{\Pr[M|A = C]\Pr[A = C]}{\Pr[M|A \neq C]\Pr[A \neq C]} = \frac{\Pr[A = C|M]}{\Pr[A \neq C|M]}. \tag{4.2}$$

Now suppose that if not for the genetic evidence there were N other suspects as likely as A to be the culprit, then it might be sensible to assume a priori

$$\frac{\Pr[A = C]}{\Pr[A \neq C]} = \frac{\frac{1}{N+1}}{\frac{N}{N+1}} = \frac{1}{N}, \tag{4.3}$$

and of course N may very well be the size of the entire population. The prosecutorial fallacy is to present the evidence as if $N = 1$ or that, a priori, the probability is .5 that the accused is the culprit. When the prosecutor implies this fallacy, as she often will, the defending attorney can counter with a more modest fallacy, and that is to ascertain the expected number of individuals K in a sufficiently large population, say of size T, such that

$$K = T \times \Pr[M|A \neq C]$$

and now assert that

$$\Pr[A = C|M] = \frac{1}{K+1}.$$

The potentially fallacious assumption here is that every other individual in the population who matches is as likely as the accused to be the culprit. Certainly where there is no other evidence other than the profile match this may be reasonable.

A weakness of this setup is the arbitrary match/nonmatch criterion and the tacit assumption

$$\Pr(M|A = C) = 1.$$

Actually,

$$\Pr(M|A = C) = 1 - \alpha$$

where α is the actual probability of a false exclusion whose properly calculated value varies with the varying laboratory criteria, joint distributional assumptions which in turn critically involve the independence of alleles both within a locus and between loci and errors other than measurement. These factors may also play an even more critical role in the estimation of

$$\Pr(M|A \neq C).$$

For a full discussion of other potential problems of this likelihood ratio, see Balding and Donnelly (1995).

5 Estimation of the Relative Frequency of a Profile

We assume first that there is a sample of profiles from a relevant population so that an estimate of this profile frequency can be made. We first describe the methods of estimation used by the three major laboratories.

For a single probe on N individuals, Lifecodes will order the $2N = n$ allelic values and for one of the profiled band values it will count all bands in the database that are within 1.8% of that value. Then that number will be divided by n which results in say $\hat{p}_1$ for band 1. The same thing is done for band 2 with result of $\hat{p}_2$. On the assumption of a heterozygote, i.e. two unequal bands, the estimate for the ith locus is $2\hat{p}_{1i}\hat{p}_{2i}$. If it is assumed that the initial band values are the same, or a homozygote, the estimate is $\hat{p}^2$, although a more conservative estimate $2\hat{p}$ is often used. This method assumes the independence of the 2 bands at the locus or in the genetic parlance the locus for

the population is assumed to be in Hardy-Weinberg equilibrium (basically the Binomial theorem). The same procedure is then applied to the several other probes making up the profile and results in a final estimate for t probes

$$\hat{P} = 2^t \prod_{i=1}^{t} \hat{p}_{1i}\hat{p}_{2i} \tag{5.1}$$

for a heterozygote. The multiplication of the probabilities for the probes assumes the independence of loci, or in genetic parlance, linkage or gametic phase equilibrium depending on the chromosomal situation. First we note that the original match is on the same gel but the "matches" in the population are on different gels. The percent standard deviation between values on different gels is close to twice (1.82) that of the percent standard deviation on the same gel, as estimated by FBI data. Hence instead of using 1.8% they should be using 3.3%.

Cellmark uses basically the same idea but uses resolution limits, one for the initial match and two for the match in the sample database.

The FBI, which uses a 5% initial match window, is a bit schizophrenic about the match in the database. It has divided the sample database for a probe arbitrarily into 31 intervals or bins and determines the bin in which the initial band falls and reports the relative frequency of that bin, Budowle et al. (1991). When the band value falls close to the boundary between 2 bins, both relative frequencies are reported and often the larger one is used. However, experts familiar with this procedure (NRC report, 1992) advise that the number in the adjacent bins be summed and the relative frequency of the total be reported. Due to the controversy on the use of DNA profiling in forensics, the National Research Council formed a committee of experts to attempt to resolve the issues. They issued the above-mentioned NRC report. In many cases the FBI has used a floating bin of $\pm 2.5\%$ and $\pm 5\%$ about the band value to be matched, neither of the two percentages are concordant with the fact that the intergel standard deviation is about 1.82 times the intragel standard deviation. This would require a $\pm 9\%$ window which the FBI does not use. Assuming that the sampling procedures are adequate and mutual independence is appropriate (both issues will be discussed subsequently), it is of interest to assess the estimate $\hat{P}$ of $P = 2^t \prod_{i=1}^{t} p_{1i}p_{2i}$, the population value. This is tantamount to t independent trinomial distributions for the heterozygotic case. Since

p_{ji} $j = 1, 2$, $i = 1, \ldots, t$ can all be considerably less than .5, it is clear that the sampling distribution of $\hat{P}$ will be skewed right. Now assuming that N individuals are all measured on t probes,

$$\text{cov } \hat{p}_{1i}\hat{p}_{2i} = E(\hat{p}_{1i}\hat{p}_{2i}) - \hat{p}_{1i}\hat{p}_{2i} = -\frac{p_{1i}p_{2i}}{n}$$

it is clear then that

$$E(\hat{p}_{1i}\hat{p}_{2i}) = p_{1i}p_{2i}\left(\frac{n-1}{n}\right),$$

which slightly underestimates each product within a probe. Since the probes are assumed independent (for the present),

$$E(\hat{P}) = \left(\frac{n-1}{n}\right)^t P$$

so the bias of underestimation becomes worse as the number of probes increase and the sample size decreases. Further and more importantly it is of interest to calculate

$$\Pr[\hat{P} \leq P].$$

This is analytically difficult so we resort to some Monte Carlo simulations to have an idea. This is presented in Table 1. Again we note that $\Pr[\hat{P} \leq P]$ can be quite large, close to 1 in fact in many cases but in addition to increasing with an increasing number of probes and decreasing sample size as the bias of $\hat{P}$ does, it also increases as the $k = 2t$ probabilities p_{ij} decrease. For the sake of ease the table is conducted using equal probabilities for each allele in each probe to obtain the values. These probabilities are then varied. This also, under the best of circumstances, puts to rest the claims of forensic laboratories that their estimates are conservative and favor the defendant. Moreover, sample databases as collected often have varying numbers of individuals on each probe. For example, Table 2 describes the number of individuals on the same probes in Cellmark's databases for three major U.S. groups, Blacks, Caucasians and Hispanics, this would require that

$$E(\hat{P}) = P\prod_{i=1}^{t}\left(\frac{n_i - 1}{n_i}\right),$$

stressing the dependence on the minimum n_i. Balding (1995), who raised the issue that the sampling distribution of $\hat{P}$ has substantial mass below P, has discussed methods for better estimation of P and has provided some shorter tables of $\Pr[\hat{P} \leq P]$ under the assumption of random samples where there is mutual independence within and between loci. We will discuss these assumptions subsequently.

6 Population Issues

Clearly the appropriate reference population is the population of identifiable possible perpetrators but this is almost always unknown. If the race or ethnicity of the culprit, who say left the crime scene profile, has been established beyond any doubt, and of the course the accused was a member of that group, then it would appear that it is the proper reference population. However, even this is a doubtful scenario. What is done in practice is to use the race (Black, Caucasian or Hispanic) of the accused to calculate the relative frequency and more generally to make the calculation for all three groups and report all of them. The NRC report (1992) strongly advised calculating a 95% upper confidence interval for each band or .1, whichever was larger, and then selecting the maximum over 3 or more major races as an interim ceiling principle that would be conservative when applying the product rule. (A more elaborate ceiling principle they suggested has never been put into practice.) Almost as soon as the interim ceiling principle was declared to always be conservative it was demonstrated by Slimowitz and Cohen (1993) that when mutual independence was not assured this principle need not be conservative, i.e. underestimate the true value. In that paper dependence was theoretically violated in the form of a mixture of differing populations or substructuring which is often the way that a genetic population is in disequilibrium.

7 Sampling Issues

In any event several populations are sampled. The major assumption that the product estimate (5.1) is based on requires a random sample of unrelated individuals that have been randomly mating for a sufficient number of generations to ensure independence. It turns out that the sampling is not random but at the laboratory's convenience and is actually referred to as "convenience" sampling, Roeder (1994). For example, Cellmark's entire Black database was obtained from a Detroit blood bank who in turn were recruiting rare donor volunteers for Black patients. The director of the blood bank indicated that they had no knowledge of how to obtain a simple random sample and no knowledge of possible familial relationships. The donors were asked for their racial category—Black,

White, etc.—but not mixed. The Caucasian database was obtained much in the same manner from the Blood Bank of Delaware. These then purport to represent U.S. Blacks and Caucasians. Further, the sample sizes themselves and their distribution among the various loci are inadequate for proper testing of independence assumptions and precise estimation, see Table 2. The FBI samples tend to be about twice as large but fall far short of what would be required to properly test their fixed binning system. This requires $496 = 31 \times 32/2$ cells in an upper triangular contingency table which is of the order of the number of individuals in their samples. In reviewing the FBI databases, it was discovered that there were about 25 apparent matches. Some of these were due to known duplicate submissions from the same individual. However, others not known to be duplicates were deleted based on a match criterion, Sullivan (1992). Again, this is a result of careless "convenience" sampling. Attempts have been made to justify small relative frequencies for a profile by looking at all possible pairs, Risch and Devlin (1992), using a criterion (actually a 2.4% window rather than a 9% window). The use of the criterion to justify the criterion can bias results.

8 Independence

The foundation for the so-called product rule rests on the assumption of mutual independence among all the alleles within and between loci. This issue was examined by Weir (1992a,b) who claimed that both the FBI's fixed bin and the floating bin approaches of Cellmark and Lifecodes indicated mutual independence and thence the propriety of the product rule. For the FBI, Weir basically admitted that a proper test of the 496 two-dimensional bins would require much larger sample sizes than had been collected aside from the fact that the samples were not randomly collected. For Hardy-Weinberg equilibrium he used intraclass correlation but this is a measure of linear association and need not have power against other types of dependence. He also applied global chi-square and likelihood ratio tests but because of the sparsity of cells, he applied bootstrap sampling methods to determine the p-values. He expected these to be powerful because they have "large degrees of freedom". That "large degrees of freedom" necessarily lead to a powerful test is a total non sequitur. In fact those tests, given the sample sizes in relation

to the number of cells, are not likely to be very powerful in detecting various forms of dependence. For the floating bin situation where bins are not determined in advance, he claimed he could not determine a global test of independence and created 10,000 profiles and tested them using a chi-square statistic calculated from the actual databases, again claiming that the results were consistent with independence. These tests would also tend to have poor power to detect various possible dependencies.

Because of the paucity of observations in relation to the potential number of pairs of bins with regard to the FBI methods or the floating bins of Lifecodes and the resolution limit of Cellmark, Geisser and Johnson (1992) devised a quantile chi-square approach to the problem of testing. The method for testing the independence of (X, Y) when the form of the exchangeable bivariate distribution is unspecified is fairly simple. Assume that a random sample (X_i, Y_i), $i = 1, \ldots, N$ has been obtained. In addition, the parental alleles are not identifiable in these databases, i.e. it is not possible to ascertain which allele of the pair is X and which is Y. Hence under independence all the values come from the same distribution. We order the $2N$ values into $Z_{(1)}, \ldots, Z_{(2N)}$ and divide them into q quantiles $Q_1, \ldots, Q_q$. We then form a $q \times q$ folded quantile table (Table 3) with sample entries n_{ij}, the number of pairs of (X_i, Y_i) that are in (Q_i, Q_j) $i \leq j$, since we can only observe $n^*_{ij} = n_{ij} + n_{ji}$, $i < j$. It was then shown that

$$Z = \frac{\left(N^{-1}\sum_1^q n_{ii} - \frac{1}{q}\right)\sqrt{N}}{\sqrt{\frac{1}{q}\left(1 - \frac{1}{q}\right)}} \to N(0,1) \tag{8.1}$$

and

$$X = \frac{q^2}{N}\sum_1^q \left(n_{ii} - \frac{N}{q^2}\right)^2 + \frac{q^2}{2N}\sum_{i>j}\left(n^*_{ij} - \frac{2N}{q^2}\right) \tag{8.2}$$

tends to χ^2 with $q(q-1)/2$ degrees of freedom. The basis for this test is that under independence $E(n_{ii}) \doteq \frac{N}{q}$ and $E(n^*_{ij}) \doteq \frac{2N}{q}$, $i = 1, \ldots, q$ and $i < j = 1, \ldots, q$.

The Z test is more useful for the substructuring alternative as a one-sided test. Since under the alternative one expects the diagonal entries to be larger than under the null hypothesis. This simple method was applied to the FBI sample databases by Geisser and Johnson (1993). This was applied to 6 different probes. For $q = 2$ it was determined that independence was rejected for D2S44, D17S79 and D1S7 at $q = 2$, and D14S13 at $q = 3$, for the Black database. For the Caucasian database, D17S79, D1S7 and D14S13 appeared

to exhibit dependence, while D2S44, D17S79, D14S13 appeared to exhibit dependence for the Hispanic database.

Recently the FBI has begun using a new probe D5S110. An analysis of this probe reveals that for Caucasians where $q = 2$, using Z the substructuring alternative $P = .05$, while for Blacks, $q = 3$ and using X, we reject at $P = .03$. The Hispanic data were divided into two groups by the FBI, Southeastern and Southwestern Hispanics. While in neither group were the tests for independence rejected at $P = .05$, in both groups there is a tendency for the diagonals to be less than expected under independence.

With regard to Cellmark, 5 probes were analyzed in a similar manner and either for $q = 2$ or 3, only MS31 was not rejected both for Caucasians and Hispanics, and only MS43 was not rejected for Blacks. So for each major database, 4 out of 5 probes exhibited dependence.

For Lifecodes, only 2 probes—D2S44 and D17S79 on a Caucasian database—were available and both exhibited dependence, Geisser and Johnson (1993).

A rigorous derivation of the theory for these tests is presented by Geisser and Johnson (1995), and also includes a quantile chi-square test for linkage equilibrium which takes advantage of the exchangeability of the alleles within a locus.

Now it is clear that the quantile chi-square test sensitivity to dependence may depend critically on q, the number of quantiles. Depending on the configuration of dependence at a locus, certain values of q may be insensitive to detecting the dependence while others may be quite sensitive. Weir (1993), in faulting the test, apparently misunderstands this issue. Only if independence is not rejected for a series of different values of q can one have some confidence that dependence is not a critical issue. Another issue he miscontrues is his assumption that the quantiles should be the binning procedure itself. The quantiles' major utility in testing independence is when a floating bin is used or when the sample sizes are inadequate for testing the FBI's fixed binning procedure.

It has also been proposed, Devlin and Risch (1993), that technical flaws in the electrophoretic process tend to make the quantile chi-square too sensitive in detecting dependence. The first problem is termed coalescence—a blurring on the autoradiograph such that presumably close but different bands are erroneously judged as the same, i.e. a homozygote. A second is termed as a null allele in which one of the band's size (or

perhaps both) is too small to appear on the autoradiograph and the band that appears is mistakenly judged as a homozygote. Further, the fact that the measurement error is highly correlated is also touted as a factor in making the test too sensitive.

Coalescence can only affect the test close to the intersection of the boundaries of the diagonal quantiles. Hence only if q is large can there be an effect if the autoradiograph is unable to discriminate between adjacent alleles. For $q = 2$ or 3 any such effect is negligible. If it is known that one of the bands is a null allele there is absolutely no effect on the test. If it is not known, the laboratory judges the two bands to be the same and is so entered into the database, which could lead to an excess along the main diagonal. All tests that have been proposed would be subject to some error in the presence of unknown null alleles. This would be confounded with the substructuring alternative. However, it is interesting to note the tendency to deficits along the main diagonal in the Southeastern and Southwestern Hispanic data for D5S110.

Since the intergel standard deviation of a band value is less than 1% of the band value, only a very small portion of the observable is subject to this measurement error correlation. This could have an effect for large q and perhaps even so only negligibly. At any rate, Devlin and Risch (1993) and Weir (1995) claim that the true alleles are independent and that any dependence that is disclosed by the test is due to the technical flaws of the electrophoretic procedure. Chakraborty et al. (1994) rehash the same arguments. The results on D5S110 reported tend to question these explanations.

But clearly even in the highly unlikely event it is in equilibrium, the observables, flawed or not, are used. If they exhibit sufficient dependence to be detected, then their use negates the product rule. It is unusual to argue that the virtues of a procedure are its technical flaws or that the best test is an insensitive one.

9 Summary

In DNA forensic profiling, the following problems and flaws are listed:

1. Technical flaws leading to coalescence and null alleles

2. False claim of favoring a defendant when the false exclusion rate appears to be orders of magnitude less than the false inclusion rate

3. Invoking or implying surreptitiously the prosecutor's fallacy

4. Lack of random sampling from well specified populations that exclude related individuals

5. Inadequate sample sizes

6. Estimates used of the relative frequency of a profile that are biased downward further belying the false claim of favoring the defendant even if all assumptions were valid

7. Reliance on the false assumption of mutual independence whether caused by substructuring or biased sampling

8. Refusal to engage in periodic, blind, external proficiency tests

9. Lack of implementation of many of the recommendations of the 1992 NRC report.

Clearly all of these problems/flaws should be resolved or corrected because DNA forensics are involved in very serious issues, mainly capital offenses and not infrequently DNA may be the only available evidence.

For some other concerns regarding the exclusion of close relatives that may also lead to a nonconservative relative frequency, see Donnelly (1992, 1994) and Balding and Donnelly (1995).

10 Acknowledgement

This work was supported in part by NIH Grant GM 25271 and the Lady Davis Trust.

References

Baird, M., Balazs, I., Giusti, A., Miyazaki, L., Nicholas, L., Wexler, K., Kanter, E., Glassberg, J., Allen, E., Rubenstein, P. and Sussman, L. (1986). Allele frequency distribution of two highly polymorphic DNA sequences in three ethnic groups and its

application to the determination of paternity. *American Journal of Human Genetics* **39** 489–501.

Balazs, I., Baird, M., Clyne, M. and Meade, E. (1989). Human population genetic studies of five hypervariable DNA loci. *American Journal of Human Genetics* **44** 182–190.

Balding, D. J. (1995). Estimating products in forensic identification. *Journal of the American Statistical Association*, to appear.

Balding, D. J. and Donnelly, P. (1995). Inference in forensic identification. *Journal of the Royal Statistical Society, A* **158** 21–53.

Berry, D. A., Evett, I. W. and Pinchin, R. (1992). Statistical inference in crime investigations using DNA profiling: single locus probes. *Journal of the Royal Statistical Society, C* **41** 499–531.

Botstein, D., White, R. L., Skolnick, M. and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* **32** 314–331.

Budowle, B., Giusti, A. M., Wayne, J. S., Baechtel, F. S., Fourney, R. M., Adams, D. E., Presley, L. A., Deadman, H. A. and Monson, K. L. (1991b). Fixed bin analysis for statistical evaluation of continuous distributions of allelic data from VNTR loci for use in forensic comparisons. *American Journal of Human Genetics* **48** 841–855.

Chakraborty, R., Zhong, Y. and Budowle, B. (1994). Non-detectability of restriction fragments and independence of DNA fragments within and between loci in RFLP typing of DNA. *American Journal of Human Genetics* **55** 391–401.

Devlin, B. and Risch, N. (1993). Physical properties of VNTR data and their impact on a test of allelic independence. *American Journal of Human Genetics* **53** 324–328.

Donnelly, P. (1992). Discussion on "Statistical inference in crime investigations using deoxyribonucleic acid profiling" by D. A. Berry, I. W. Evett and R. Pinchin. *Applied Statistics* **41** 524–525.

Donnelly, P. (1995). The non-independence of matches at different loci in single-locus DNA profiles. *Heredity*, to appear.

Geisser, S. and Johnson, W. (1992). Testing Hardy-Weinberg equilibrium on allelic data from VNTR loci. *American Journal of Human Genetics* **51** 1084–1088.

Geisser, S. and Johnson, W. (1993). Testing independence of fragment lengths within VNTR loci. *American Journal of Human Genetics* **53** 1103–1106.

Geisser, S. and Johnson, W. (1995). Testing independence when the form of the bivariate distribution is unspecified. *Statistics in Medicine* (to appear).

Jeffreys, A. J., Wilson, V. and Thein, S. L. (1985a). Hypervariable "minisatellite" regions in human DNA. *Nature* **314** 67–73.

Jeffreys, A. J., Wilson, V. and Thein, S. L. (1985b). Individual-specific "fingerprints" of human DNA. *Nature* **316** 76–79.

National Research Council (1992). *DNA Technology in Forensic Science*. National Academy Press, Washington, D.C.

Risch, N. J. and Devlin, B. (1992). On the probability of matching DNA fingerprints. *Science* **225** 717–720.

Roeder, K. (1994). DNA fingerprinting: A review of the controversy. *Statistical Science* **9** 222–278.

Slimowitz, J. R. and Cohen, J. E. (1993). Violations of the ceiling principle: Exact conditions and statistical evidence. *American Journal of Human Genetics* **53** 314–323.

Sullivan, P. (1992). DNA fingerprint matches. *Science* **256** 1743–1744.

Thompson, W. C. (1993). Evaluating the admissibility of new genetic identification tests: Lessons from the "DNA" war. *Journal of Criminal Law and Criminology* **84** 1 22–104.

Thompson, W. C. and Schumann, E. L. (1987). Interpretation of statistical evidence in criminal trials. *Law and Human Behavior* **11** 167–187.

Weir, B. S. (1992a). Independence of VNTR alleles defined by fixed bins. *Genetics* **130** 873–887.

Weir, B. S. (1992b). Independence of VNTR alleles defined as floating bins. *American Journal of Human Genetics* **51** 992–997.

Weir, B. S. (1993). Tests for independence of VNTR alleles defined as quantile bins. *American Journal of Human Genetics* **53** 1107–1113.

Weir, B. S. (1994). Comment on DNA fingerprinting. *Statistical Science* **9** 266–267.

Table 1: Probability that the product estimate is less than the true product, based on 30,000 simulations

	k = 2	k = 4	k = 6	k = 8	k = 10
p = .01					
n					
5	0.9977667	1.0000000	1.0000000	1.0000000	1.0000000
10	0.9909333	1.0000000	1.0000000	1.0000000	1.0000000
20	0.9705667	0.9989667	1.0000000	1.0000000	1.0000000
50	0.8452000	0.9762333	0.9964000	0.9992000	0.9998000
100	0.6021000	0.8405000	0.9351000	0.9741333	0.9895000
200	0.5702333	0.7006000	0.7842667	0.8447667	0.8821333
400	0.5413000	0.6120000	0.6496667	0.6843333	0.7119000
p = .02					
n					
5	0.9921000	0.9999000	1.0000000	1.0000000	1.00000000
10	0.9691000	0.9993333	1.0000000	1.0000000	1.0000000
20	0.8936333	0.9884667	0.9987667	0.9999000	1.0000000
50	0.5983000	0.8382667	0.9375000	0.9741667	0.9891667
100	0.5693667	0.7021667	0.7887000	0.8467667	0.8849333
200	0.5400667	0.6231000	0.6582333	0.6914667	0.7172000
400	0.5355667	0.5808667	0.6105333	0.6316667	0.6526667
p = .05					
n					
5	0.9572000	0.9981000	0.9999333	1.0000000	1.0000000
10	0.8488667	0.9771333	0.9964000	0.9994000	0.9998667
20	0.7391667	0.8550333	0.9355667	0.9740333	0.9890667
50	0.6603333	0.7026000	0.7631000	0.8041667	0.8385000
100	0.5868333	0.6015333	0.6425333	0.6743333	0.6979000
200	0.5587333	0.5709667	0.5979000	0.6198667	0.6371333
400	0.5399000	0.5487667	0.5674333	0.5814333	0.5949667
p = .1					
n					
5	0.8511000	0.9777667	0.9964667	0.9995333	0.9999333
10	0.7394333	0.8522667	0.9332333	0.9733333	0.9891000
20	0.7011000	0.7396667	0.7856667	0.8406667	0.8760667
50	0.5885667	0.6008333	0.6486667	0.6829667	0.7041000
100	0.5578333	0.5680667	0.6026333	0.6221667	0.6405333
200	0.5456667	0.5578333	0.5734000	0.5877667	0.6044000
400	0.5355667	0.5375667	0.5511333	0.5602667	0.5712333
p = .2					
n					
5	0.7477333	0.8454667	0.9265333	0.9670333	0.9865333
10	0.7099333	0.7322000	0.7863333	0.8344333	0.8721000
20	0.6222333	0.6546333	0.6880667	0.7220000	0.7476333
50	0.5778333	0.5902333	0.6156667	0.6370333	0.6563000
100	0.5549000	0.5624000	0.5769333	0.5921333	0.6048333
200	0.5359000	0.5431667	0.5553000	0.5696333	0.5804000
400	0.5237333	0.5257000	0.5360667	0.5433333	0.5505000

Table 2: Number of individuals with values on Cellmark probes

1. Pairwise

Pairs	Blacks	Caucasians	Hispanics
MS1/G3	10	235	155
MS1/YNH24	91	154	104
MS1/MS43	128	177	178
MS1/MS31	155	210	154
MS31/YNH24	80	110	94
MS31/MS43	103	189	151
MS31/G3	10	171	133
MS43/YNH24	65	146	95
MS43/G3	10	253	142
YNH24/G3	16	154	93

2. Omit One Probe

Omitted Probe			
MS1	2	79	63
MS31	2	108	73
MS43	2	77	72
G3	31	91	76
YNH24	8	153	109

3. None Omitted

	2	75	59
Total on each Probe			
MS1	240	262	215
MS31	238	264	183
MS43	223	294	192
G3	200	325	168
YNH24	146	208	110

Table 3: Folded Quantile Contingency Table

	Q_1	Q_2		Q_q
Q_1	n_{11}	n^*_{12}	$\cdots$	n^*_{1q}
Q_2		n_{22}	$\cdots$	n^*_{2q}
			$\ddots$	$\vdots$
Q_q				n_{qq}